DOCUMENT RESUME

ED 408 306 TM 026 508

AUTHOR Kim, Seock-Ho; Cohen, Allan S.
TITLE A Comparison of Linking and Concurrent Calibration under the Graded Response Model.
PUB DATE Mar 97
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 1997).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; Comparative Analysis; *Computer Assisted Testing; *Equated Scores; Estimation (Mathematics); Item Bias; *Item Response Theory; Testing Problems
IDENTIFIERS *Calibration; *Graded Response Model; Item Discrimination (Tests); Linking Metrics

ABSTRACT

Applications of item response theory to practical testing problems including equating, differential item functioning, and computerized adaptive testing, require that item parameter estimates be placed onto a common metric. In this study, two methods for developing a common metric for the graded response model under item response theory were compared: (1) linking separate calibration runs using equating coefficients from the characteristic curve method; and (2) concurrent calibration using the combined data of the base and target groups. Concurrent calibration yielded consistently albeit only slightly smaller root mean square differences for both item discrimination and location parameters. Similar results were observed for Euclidean distances between estimates and parameters. (Contains 2 figures, 6 tables, and 24 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 408 306

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
SEOCK-HO KIM
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

A Comparison of Linking and Concurrent Calibration Under the Graded Response Model

Seock-Ho Kim
The University of Georgia
Allan S. Cohen
University of Wisconsin–Madison

March, 1997
Running Head: LINKING AND CONCURRENT CALIBRATION

Paper presented at the annual meeting of the American Educational Research Association, Chicago.

TM026508

BEST COPY AVAILABLE

A Comparison of Linking and Concurrent Calibration Under the Graded Response Model

Abstract

Applications of item response theory to practical testing problems including equating, differential item functioning, and computerized adaptive testing, require item parameter estimates be placed onto a common metric. In this study, we compared two methods for developing a common metric for the graded response model under item response theory: (1) linking separate calibration runs using equating coefficients from the characteristic curve method and (2) concurrent calibration using the combined data of the base and target groups. Concurrent calibration yielded consistently albeit only slightly smaller root mean square differences for both item discrimination and location parameters. Similar results were observed for Euclidean distances between estimates and parameters.

Key words: concurrent calibration, equating, graded response model, item response theory, linking, MULTILOG.

Introduction

Studies of horizontal and vertical equating and studies of differential item functioning under item response theory (IRT) require that item parameters from two or more data sets be expressed on a common metric. In this paper, we refer to linking as developing a common metric in IRT by transforming a set of item parameter estimates from one metric onto another, base metric. It is also possible, under IRT, to develop a common metric by simultaneously calibrating a combined data set. In spite of the importance of the metric of the θ scale under IRT, however, very little work has directly addressed the issues of linking versus concurrent calibration or the issue of the identification problem for the graded response model. Previous research on these issues, in fact, (Kim & Cohen, in press; Peterson, Cook, & Stocking, 1983; Wingersky, Cook, & Eignor, 1986) has focused solely on dichotomous IRT models.

This is unfortunate, as the use of IRT models for polytomously-scored items is increasing in popularity, due largely to the widespread use of constructed-response format items particularly in the context of performance assessment. As is the case for dichotomous IRT models, successful applications of IRT with polytomous models depends upon the metric of item and ability parameters. In this study, we compare linking and concurrent calibration methods used for developing a common ability metric under Samejima's (1969, 1972) graded response model.

Under Samejima's (1969, 1972) graded response model, the category response function $P_{jk}(\theta)$ is the probability of response k to item j as a function of θ. This function is defined as

$$P_{jk}(\theta) = \begin{cases} 1 - P^*_{j1}(\theta) & \text{when } k = 1 \\ P^*_{j(K-1)}(\theta) & \text{when } k = K \\ P^*_{j(k-1)}(\theta) - P^*_{jk}(\theta) & \text{otherwise,} \end{cases} \quad (1)$$

where $P_{jk}^{*}(\theta)$is the boundary response function in the form of the logistic model given by

$$P_{jk}^{*}(\theta) = \{1 + \exp[-\alpha_j(\theta - \beta_{jk})]\}^{-1}, \tag{2}$$

where α_j is the discrimination parameter for item j, β_{jk} is the location parameter, and θ is the trait level parameter. With $P_{j0}^{*}(\theta) = 1$ and $P_{jK}^{*} = 0$, the category response function can be succinctly written as

$$P_{jk}(\theta) = P_{j(k-1)}(\theta) - P_{jk}^{*}(\theta), \tag{3}$$

where $k = 1(1)K$ and K is the total number of categories. Figures 1 and 2 illustrate the category response functions and the boundary response functions, respectively, for a typical graded response model item with five ordered response categories: $\alpha_j = 1.46$, $\beta_{j1} = -.35$, $\beta_{j2} = .67$, $\beta_{j3} = .97$, $\beta_{j4} = 1.94$.

Insert Figures 1 and 2 about here

The purpose of equating is to convert test scores obtained from the metric of one test to the metric of a second test. In horizontal equating, the tests to be equated are at the same level of difficulty and the ability distributions of examinees are comparable. Horizontal equating is required when multiple forms of a test are needed. In vertical equating, the tests to be equated are at different levels of difficulty and the ability distributions of examinees are not comparable. Vertical equating is required so that a single scale can be used to make comparisons of abilities of examinees at different levels (e.g., different grade levels or different age groups). Under IRT, equating may not be necessary, if item parameters from the two tests are on the same metric. Hence, in IRT the task of equating is reduced to that of developing a common metric.

Both equating of test scores from various tests and linking of item parameters can be carried out under several different designs (Vale, 1986). In this paper, we consider the anchor test design in which two tests contain a set of common items and the tests are administered to two groups of examinees either with comparable or different ability levels.

When separate calibration runs are used for dichotomously scored IRT models, three classes of linking methods are available for obtaining the linking or equating coefficients, A and B: characteristic curve methods (Divgi, 1980; Haebara, 1980; Stocking & Lord, 1983), the minimum chi-square method (Divgi, 1985), and mean and sigma methods (Linn, Levine, Hasting, & Wardrop, 1981; Loyd & Hoover, 1980; Marco, 1977; Stocking & Lord, 1983). Each method has been extended to the graded response model by Baker (1992), Kim and Cohen (1995), and Cohen and Kim (1993), respectively. The transformation coefficients are obtained from the item parameter estimates of the common items on the two tests. In general, if there are two sets of item parameter estimates, one from the base group and the other from the target group, the task is to place item and ability estimates of the target group onto the metric of the base group. Item parameter estimates from the target group, including those for the common items, are placed onto the metric of the base group via the coefficients A and B. After the metric transformation and in order to achieve symmetry of transformation, the item parameter estimates from the base group and the transformed item parameter estimates from the target group for the common items can be averaged to obtain the final estimates (Hambleton & Swaminathan, 1985).

Concurrent calibration involves estimating item and ability parameters simultaneously, typically on a single computer run. This is done by combining data from both (or several) groups and treating items not taken by a

particular group as not reached or missing (Lord, 1980). A variation of this is also possible in which the parameter estimates of the common items from the base group are set to be fixed and the remaining item parameters are estimated using data from the target group. Note, in either case with concurrent calibration, there will be only one set of parameter estimates for the common items.

Concurrent calibration of the graded response model is presently possible using marginal maximum likelihood estimation (MMLE) as implemented, for example, in the computer program MULTILOG (Thissen, 1991). In MMLE (e.g., Bock & Aitkin, 1981), the joint likelihood is marginalized under the assumption that a population distribution exists. When there are two groups of examinees, MULTILOG default options calibrate items by constructing a unit normal metric for ability parameters of the base group. The mean ability of the target group is then obtained empirically along with the item parameters while fixing the standard deviation at unity. MULTILOG default options can also be overridden so that the mean and the standard deviation of the target group can be specified differently. If the target group population distribution of ability is truly different from that of the base group, then marginalization of the likelihood function should be performed using two different ability distributions.

One unresolved issue in the context of concurrent calibration under MMLE for the graded response model, is the effect of the form of the population ability distribution. In addition, there is a possible concern regarding the specification of the target group population parameters. In a horizontal equating situation, this specification may not cause serious problem, as the two distributions of abilities are generally comparable and the difficulty level of a well-designed test is typically matched to the ability

of the examinee groups. In a vertical equating situation, however, the specification becomes somewhat more complicated, particularly if the two ability distributions differ not only in location but also in variability.

In part, concurrent calibration can potentially remove some equating errors, which might arise in the case of linking, due to using results from the two separate calibration runs. It could possibly also remove some of the arbitrariness of the decisions made in linking. Concurrent calibration, however, may not always be either possible or economical. For example, item parameter estimates obtained on earlier forms of a test will generally differ to some extent from current estimates. Subsequent combination of existing data with new data just to achieve concurrent calibration results may also incur different equating errors.

Comparative studies of differences in the metrics obtained from linking and concurrent calibration have not been reported with respect to the graded response model. In the present study, therefore, we focus on this issue in the context of a recovery study.

Method

Data Generation

Data for the simulation study were generated for a 30-item test under the graded response model using the computer program GENIRV (Baker, 1988). The item parameters used to generate the data (see Table 1) were based on calibration results of the mathematics tests developed as a part of the Wisconsin Student Assessment System (Webb, 1994). All items had five ordered categories. Note that the mean of the location parameters β_{jk} was .962 and the standard deviation was .893.

Insert Table 1 about here

In a typical linking or concurrent calibration situation, there are two groups of examinees, the base and the target groups. In this study the following three different sample size combinations of a base group with a target group were used: (1) a base group with 300 examinees and a target group with 300 examinees (we denote this Base 300/Target 300, respectively), (2) Base 1,000/Target 1,000, and (3) Base 1,000/Target 300. The sample size of 300 was used to simulate a small sample. Previous research on the graded response model (Reise & Yu, 1990) indicated that at least 500 examinees were needed to achieve an adequate calibration.

Using these guidelines, a sample of 300 would be considered a small sample.

The ability of the base group was generated normal with a mean of 1 and a standard deviation of 1 [i.e., N(1,1)]. This set of generating ability parameters were used so that the base group's ability essentially matched the difficulty of the test. There were two different target group ability distributions generated: N(0,1) and N(1,1). The N(0,1) target group was generated to have a group lower in ability than the base group and also so the test would be hard for this group. This base group–target group combination simulated a vertical equating situation. To simulate a horizontal equating situation, both the base and target groups were generated to have the same N(1,1) ability distribution, one that was also matched to the difficulty of the test.

Data were first generated for the Base 300/Target 300 and Base 1,000/Target 1,000 combinations for both horizontal and vertical simulations. To sim-

ulate a large group–small group equating situation, a Base 1,000 sample was randomly paired with a Target 300 sample. This was done for both the vertical and horizontal equating situations. 100 replications were generated for each of the three sample size by two ability group conditions.

Number of Common Items and Item Parameter Estimation

For each combination of a target group and base group, three different lengths of common items sets were used: 5, 10, and 30 items. For the 5-common item condition, items 1–5 in Table 1 were used. For the 10-common item condition, items 1–10 were used. The 30-common item condition simulated a typical differential item functioning detection situation in which all of the items need to be placed onto the same metric before comparisons could be made.

The computer program MULTILOG was used to estimate the item parameters for the separate calibration runs followed by linking. Default MULTILOG options under the graded response model were used for these calibrations. First, base group and target group item parameters were estimated separately. Next, the test characteristic curve method for linking under the graded response model (Baker, 1992), as implemented in the computer program EQUATE (Baker, 1993), was used to obtain the transformation coefficients A and B. These coefficients were used to link the target metric to the base group metric. The transformation equations are

$$a^*_{j\mathrm{T}} = a_{j\mathrm{T}}/A \tag{4}$$

and

$$b^*_{jk\mathrm{T}} = A \times b_{jk\mathrm{T}} + B, \tag{5}$$

where * indicates the values on the base group metric and the subscript T designates the estimate is from the target group.

Since we had three different linking situations corresponding to the three lengths of common item sets, for each combination of the base group and a target group, three EQUATE runs were performed. In case of the 5-common item condition, for example, the EQUATE run produced linking coefficients A and B based on these 5 items. Then, using A and B, item parameter estimates from the target group were placed onto the metric of the base group. Finally, the item parameter estimates from the common items were averaged to obtain the linked item parameter estimates, as recommended by Hambleton and Swaminathan (1985). For the 5-common item condition, this resulted in estimates of item parameters for 55 items after the linking. Similarly, for the 10-common item condition, there were estimates of item parameters for 50 items after linking, and for the 30-common item condition, there were 30 estimates for items after linking. A total of 1,800 EQUATE runs were performed, that is, 100 replications for the three EQUATE runs of the base group and the N(0,1) target group as well as the N(1,1) target group in each of the three combinations of sample sizes.

For the concurrent calibrations, the combined data for the base and target groups were used. A single combined data set was analyzed three times using MULTILOG, once for each of the three common item conditions. Altogether, 1,800 MULTILOG calibration runs were performed. For the MULTILOG concurrent calibration runs, all program default options were used resulting in MMLE of item parameters under the graded response model.

Equating and Evaluation Criteria

The final estimates for item parameters from linking were all expressed on the metric of the base group. The parameter estimates are in fact not based on an empirical ability metric but rather on a posterior metric obtained from the marginalization process. These final estimates from linking and concurrent calibrations, therefore, may not be directly comparable.

In order to make comparisons of the estimates, additional EQUATE runs were performed to place all item parameter estimates onto the metric of generating item parameters. In the case of the 5-common item condition, 55 items were equated to the metric of generated item parameters. For the 10-common item condition, 50 items were equated back to the metric of the generated item parameters. All together, 3,600 additional EQUATE runs were required to place the final estimates from the concurrent calibration runs onto a common metric of the generating parameters.

One means of evaluating results from the different methods of obtaining a common metric is to compare equating coefficients to expected values. In the separate calibration case, the equating coefficients were compared with the theoretically expected values. For concurrent calibration, linking coefficients are not available.

Instead, we compared the ability parameter estimates of the target group with the expected values.

A more definitive description is possible, however, in a recovery study. Since it is possible that a method of obtaining a common metric may function better at recovery of one type of item parameter than another, root mean square differences (RMSDs) between the estimates and the generating parameters can be used to provide an indication of the quality of the recovery

and, thereby, of the quality of linking versus concurrent calibration. The smaller the RMSDs, the better the method is in recovering the underlying metric. RMSDs were calculated separately for each parameter, once for the item discrimination parameter and once for the set of item location parameters. The RMSD for item discrimination is defined as

$$\sqrt{\frac{1}{n}\sum_{j=1}^{n}(a_j - \alpha_j)^2}, \tag{6}$$

where n is the total number of items. Recall that the total number of items were 55, 50, and 30 for each common item condition of 5, 10, and 30 items, respectively. For item location parameters, the RMSD is defined as

$$\sqrt{\frac{1}{4n}\sum_{j=1}^{n}\sum_{k=1}^{4}(b_{jk} - \beta_{jk})^2}. \tag{7}$$

Note that the item parameter estimates for both linking and concurrent calibration were equated back to the metric of the generating item parameters before calculating the RMSDs.

It is also useful to consider a single index which can simultaneously describe the quality of the recovery for all item parameters. The mean Euclidean distance (MED) provides such an index. The MED is the average of the square roots of the sum of the squared differences between the discrimination and difficulty parameter estimates and their generating values. The MED is defined as

$$\frac{1}{n}\sum_{j=1}^{n}\sqrt{(\hat{\xi}_j - \xi_j)'(\hat{\xi}_j - \xi_j)}, \tag{8}$$

where $\hat{\xi}_j = (a_j, b_{j1}, \ldots, b_{j4})'$ and $\xi_j = (\alpha_j, \beta_{j1}, \ldots, \beta_{j4})'$. MEDs were calculated between the underlying parameters and their estimates. One

caveat in using the MED, of course, is that item discrimination and location parameters are not expressed in comparable and interchangeable metrics. Even so, the MED does provide a potentially useful descriptive index.

Results

Linking Coefficients and Population Parameter Estimates

For separate calibration/linking results, the theoretically expected values of A and B for placing the N(0,1) target group metric onto the base group metric, which was generated as N(1,1), are 1 and -1, respectively. For placing the N(1,1) target group onto the N(1,1) base group metric, the expected values of A and B are 1 and 0, respectively. Summary statistics of the equating coefficients for the two different sample sizes from the separate calibration runs for the two different target group ability distributions by three numbers of common items are reported in Table 2.

Insert Table 2 about here

For the N(1,1) target group, differences in equating coefficients from expected values were quite small for all simulated conditions. The A and B were essentially 1 and 0 for all common item conditions. For the N(1,0) target group, however, the A and B were not close to the theoretically expected values. The A and B were approximately 1 and $-.57$, respectively. It is interesting to observe that, based on the sizes of the standard deviations, the values of A and B were very consistent across all replications. There were no clear effects on the values of A and B due to sample sizes or numbers of the common items. But, what happened for the N(0,1) target group is that

MULTILOG yielded location parameter estimates that were shifted toward the ability distribution used in the marginalization.

For concurrent calibration, MULTILOG set the base group's ability metric to N(0,1). The mean ability of the target group (i.e., the population parameter also called the hyperparameter) was jointly estimated along with the item parameters. Standard deviations of ability for the base group and the target group were both fixed at 1. Since we used the base group of N(1,1), the expected population mean was -1 for the N(0,1) target group and 0 for the N(1,1) target group.

Insert Table 3 about here

Table 3 contains means and standard deviations of the population parameter estimates from concurrent calibrations over 100 replications for different sample sizes, target group ability conditions, and three common item conditions. As can be seen in Table 3, the posterior population means of the target group N(0,1) were very close to the expected value. All values were negligibly smaller than the expected value of -1. The mean hyperparameters for the N(1,1) target group were also very close to the expected value 0. These results suggest that the underlying population parameters were recovered very well in both ability group conditions.

Root Mean Square Differences

Recovery of the underlying parameters can be more precisely evaluated with RMSDs between the transformed estimates and the generating parameters. The results for item discrimination, summarized in Tables 4 and 5, indicate that concurrent calibration consistently yielded smaller RMSDs for item

discrimination across all conditions, although these differences all appear primarily in the third decimal place.

Insert Tables 4 and 5 about here

RMSDs for item discrimination for the N(1,1) target group were smaller than for the N(0,1) target group. As can be seen in Table 6, there was a clear tendency for the RMSDs for item discrimination to decrease as the number of common items increased. This was particularly the case for the 30-common item simulations.

RMSDs for item location parameters are also reported in Tables 4 and 5 for separate calibration and concurrent calibration, respectively. There did not appear to be any systematic relationship between the distribution of the target group's ability and the size of RMSDs for item difficulty. As the number of common items increased, however, the size of the RMSDs for item location parameters decreased.

Mean Euclidean Distances

Trends for MEDs between item parameter estimates and underlying parameters were similar to those reported for RMSDs. Table 6 presents the MED results. Concurrent calibration consistently yielded very slightly smaller MEDs for all conditions than did separate calibration/linking. As was noted for RMSDs, however, such differences were primarily in the third decimal place. The size of the average MEDs was found to decrease, however, as the number of common items increased. Also, for both separate calibration and concurrent calibration, the N(1,1) target ability condition yielded slightly smaller MEDs than did the N(0,1) target ability condition.

Insert Table 6 about here

Summary and Discussion

The comparability of IRT item parameter estimates across different tests measuring the same underlying ability is an important concern for test developers and researchers since all decisions about examinees are derived from these estimates. Very few studies on this topic have focused on the graded response model. This is indeed unfortunate given the recent upsurge in interest in performance assessment. A number of different methods are available for developing common metrics for the graded response model, but they do not all yield the same ability estimates. Which method to choose is often a matter of uncertainty and concern. In this paper, we have presented simulation results using two methods for obtaining a common metric under the graded response model. The two methods were linking of separately calibrated metrics using linear equating coefficients A and B obtained from the test characteristic curve method and concurrent calibration of the combined data. Both methods were simulated using MMLE.

The recovery study approach permitted comparisons to be made of the similarities between generating parameters and item parameter estimates obtained after transformation of the results to the underlying metric. The simulation results indicated that recovery via concurrent calibration was consistently, albeit only slightly better than recovery from separate calibration and linking. Note that this result was not fully consistent with the result from a previous study under the dichotomous IRT model (cf., Kim & Cohen, in press). In that study, separate calibration followed by linking

yielded better results for the small number of common items.

Differences between the methods compared in this study were primarily ones inherent to the indeterminacy of the IRT ability metric. As is well-known, the ability metric in IRT is unique up to a linear transformation. Both linking and concurrent calibration are closely related to the problem of the metric indeterminacy. Computer programs for estimating item and ability parameters under IRT resolve this problem in different ways. For the MULTILOG runs in this study, ability parameters were not estimated. Instead, the underlying metric provided by MULTILOG was the normalized posterior of the base group's latent ability distribution. One of the factors playing a role in determining the metric, therefore, is the form of the prior ability distribution used in marginalization.

The scales resulting from the two different methods were also not the same. Therefore, before RMSDs and MEDs could be obtained, it was necessary to perform an additional linking for both the linked item parameter estimates from separate calibration/linking and the parameter estimates from the concurrent calibration results in order to place these results on the underlying metric. A linear transformation, such as the one used in this study due to Stocking and Lord (1983), can be used to put item parameter estimates onto the metric of underlying parameters. Remaining differences between estimates and parameters are generally due to estimation errors.

The averaging procedure recommended by Hambleton and Swaminathan (1985) is only one of many possible ways to achieve symmetry in transformation of parameter estimates to a common scale. In general, such methods have not been widely studied and, in fact, may not be appropriate in all cases (e.g., linking new items into an existing bank or detection of differential item functioning).

Results from the present study suggest that, in general, concurrent calibration results differed slightly from those from separate calibration/linking. When the ability distributions of the base and target groups were well matched to the distribution of item location parameters, however, small errors were found for both methods. Increasing the number of common items also served to decrease the size of errors. Further studies of methods for obtaining a common metric under the graded response model would be useful. In particular, it would be interesting to investigate the impact of the form of prior population distribution used under concurrent calibration in MMLE.

References

Baker, F. B. (1988). *GENIRV: A program to generate item response vectors* [Computer program]. Madison, University of Wisconsin, Department of Educational Psychology, Laboratory of Experimental Design.

Baker, F. B. (1992). Equating tests under the graded response model. *Applied Psychological Measurement, 16,* 87–96.

Baker, F. B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46,* 443–459.

Cohen, A. S., & Kim, S.-H. (1993, April). *A comparison of equating methods under the graded response model.* Paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, GA.

Divgi, D. R. (1980, April). *Evaluation of scales for multilevel test batteries.* Paper presented at the annual meeting of the American Educational Research Association, Boston.

Divgi, D. R. (1985). A minimum chi-square method for developing a common metric in item response theory. *Applied Psychological Measurement, 9,* 413–415.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research, 22*, 144–149.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Kim, S.-H., & Cohen, A. S. (1995). A minimum χ^2 method for equating tests under the graded response model. *Applied Psychological Measurement, 19*, 167–176.

Kim, S.-H., & Cohen, A. S. (in press). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*.

Linn, R. L., Levine, M. V., Hasting, C. N., & Wardrop, J. L. (1981). An investigation of item bias in a test of reading comparison. *Applied Psychological Measurement, 5*, 159–173.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement, 17*, 169–194.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139–160.

Petersen, N. S., Cook, L. L., & Stocking, M. L. (1983). IRT versus conventional equating methods: A comparative study of scale stability. *Journal of Educational Statistics, 8*, 137–156.

Reise, S. P., & Yu, J. (1990). Parameter recovery in the graded response model using MULTILOG. *Journal of Educational Measurement, 27*, 133–144.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Samejima, F. (1972). A general model for free response data. *Psychometrika Monograph Supplement*, No. 18.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201–210.

Thissen, D. (1991). *MULTILOG user's guide: Multiple, categorical item analysis and test scoring using item response theory* [Computer program]. Chicago: Scientific Software.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333–344.

Webb, N. L. (1994). *Wisconsin performance assessment development project: Analysis and technical report for fiscal year 1993-94*. Madison: University of Wisconsin, Wisconsin Center for Educational Research.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1986, April). *Specifying the characteristics of linking items used for item response theory item calibration*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Table 1
Generating Item Parameters

Item	Parameters				
	α_j	β_{j1}	β_{j2}	β_{j3}	β_{j4}
1	1.46	−.35	.67	.97	1.94
2	1.73	.18	.90	1.29	1.94
3	1.81	−.37	.03	.91	2.29
4	1.53	−.56	−.13	.80	2.22
5	1.57	−.38	.49	1.04	2.33
6	1.89	−.61	.63	1.37	2.34
7	1.89	.01	.67	1.33	2.18
8	1.84	−.23	.31	.98	2.46
9	1.93	−.31	.60	1.27	2.44
10	2.53	−.36	.53	1.20	2.34
11	1.79	−.52	.39	1.54	2.00
12	1.86	−.53	−.12	1.27	2.25
13	2.35	.06	.99	1.50	2.20
14	1.79	−.20	.49	1.00	2.40
15	2.12	.20	.56	1.40	2.00
16	2.07	−.44	.18	1.34	2.15
17	2.19	−.01	.39	1.36	2.01
18	2.40	.10	1.06	1.61	2.01
19	1.79	−.10	.35	1.01	2.22
20	2.12	.19	1.10	1.45	2.01
21	1.75	−.57	.93	1.31	2.01
22	2.16	.59	.91	1.32	2.01
23	1.86	−.02	.63	1.28	2.01
24	2.22	.52	.85	1.43	2.01
25	2.18	−.27	.58	1.24	2.25
26	2.01	−.66	.41	1.63	2.24
27	2.14	.05	.71	1.03	2.09
28	2.13	.43	1.15	1.47	2.06
29	2.12	.08	.70	1.12	2.09
30	2.05	.19	.61	.94	2.38

Table 2

Mean and Standard Deviation of Equating Coefficients over 100 Replications

Sample Size		Target		Coefficient *A*		Coefficient *B*	
Base	Target	Ability[a]	NC[b]	Mean	SD	Mean	SD
300	300	N(0,1)	5	1.040	.008	−.580	.010
			10	1.040	.005	−.573	.007
			30	1.041	.004	−.570	.003
		N(1,1)	5	.993	.009	−.013	.008
			10	.997	.006	−.009	.006
			30	1.001	.005	−.006	.002
1000	1000	N(0,1)	5	1.050	.004	−.576	.003
			10	1.047	.002	−.576	.002
			30	1.044	.001	−.572	.002
		N(1,1)	5	1.003	.004	−.004	.001
			10	1.002	.002	−.002	.001
			30	1.001	.001	.001	.001
1000	300	N(0,1)	5	1.045	.003	−.577	.005
			10	1.044	.002	−.571	.003
			30	1.040	.001	−.568	.002
		N(1,1)	5	.998	.005	−.008	.005
			10	1.001	.003	−.005	.004
			30	1.000	.002	−.004	.002

[a]Base ability is N(1,1).

[b]Number of Common Items

Table 3

Mean and Standard Deviation of Population $\hat{\mu}$ over 100 Replications

Sample Size		Target		Population $\hat{\mu}$	
Base	Target	Ability[a]	NC[b]	Mean	SD
300	300	N(0,1)	5	−1.046	.039
			10	−1.035	.030
			30	−1.029	.024
		N(1,1)	5	.005	.025
			10	.005	.022
			30	.004	.018
1000	1000	N(0,1)	5	−1.046	.017
			10	−1.039	.014
			30	−1.032	.012
		N(1,1)	5	.004	.013
			10	.004	.012
			30	.005	.009
1000	300	N(0,1)	5	−1.037	.030
			10	−1.023	.023
			30	−1.017	.017
		N(1,1)	5	.005	.022
			10	.005	.018
			30	.004	.013

[a]Base ability is N(1,1).

[b]Number of Common Items

Table 4

Mean and Standard Deviation of Root Mean Square Differences over 100 Replications from Separate Calibration

Sample Size		Target		α_j		β_{j1}		β_{j2}		β_{j3}		β_{j4}	
Base	Target	Ability[a]	NC[b]	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
300	300	N(0,1)	5	.184	.020	.091	.010	.081	.008	.094	.010	.147	.022
			10	.176	.019	.085	.009	.077	.008	.090	.009	.136	.017
			30	.128	.015	.065	.009	.058	.008	.067	.009	.105	.015
		N(1,1)	5	.177	.020	.103	.010	.079	.009	.074	.008	.106	.012
			10	.169	.019	.097	.010	.076	.008	.071	.007	.099	.010
			30	.124	.015	.074	.011	.057	.008	.054	.007	.075	.011
1000	1000	N(0,1)	5	.100	.012	.050	.006	.044	.005	.052	.007	.081	.011
			10	.095	.012	.047	.005	.041	.004	.049	.006	.074	.009
			30	.068	.009	.035	.004	.031	.004	.036	.005	.055	.008
		N(1,1)	5	.095	.010	.058	.006	.042	.005	.042	.004	.058	.008
			10	.091	.010	.054	.006	.040	.005	.039	.004	.053	.006
			30	.064	.008	.041	.006	.030	.004	.029	.003	.040	.005
1000	300	N(0,1)	5	.152	.019	.068	.008	.066	.008	.084	.011	.134	.022
			10	.145	.019	.064	.007	.062	.008	.080	.010	.123	.018
			30	.104	.017	.048	.007	.046	.006	.058	.008	.094	.015
		N(1,1)	5	.144	.018	.084	.009	.063	.007	.060	.007	.086	.011
			10	.137	.017	.080	.009	.061	.007	.057	.006	.080	.009
			30	.098	.014	.060	.008	.046	.006	.043	.005	.062	.009

[a]Base ability is N(1,1).

[b]Number of Common Items

Table 5

Mean and Standard Deviation of Root Mean Square Differences over 100 Replications from Concurrent Calibration

Sample Size		Target		α_j		β_{j1}		β_{j2}		β_{j3}		β_{j4}	
Base	Target	Ability[a]	NC[b]	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
300	300	N(0,1)	5	.181	.019	.089	.012	.079	.009	.092	.010	.143	.022
			10	.174	.019	.084	.009	.076	.008	.088	.009	.132	.017
			30	.118	.013	.062	.009	.055	.007	.061	.008	.090	.012
		N(1,1)	5	.173	.018	.101	.010	.077	.008	.073	.007	.104	.011
			10	.168	.019	.096	.010	.075	.008	.071	.007	.098	.010
			30	.122	.015	.073	.011	.057	.008	.053	.007	.074	.011
1000	1000	N(0,1)	5	.098	.011	.050	.006	.044	.004	.050	.005	.078	.008
			10	.094	.012	.047	.005	.041	.004	.048	.005	.072	.008
			30	.064	.008	.033	.004	.029	.004	.033	.004	.047	.006
		N(1,1)	5	.093	.010	.056	.006	.041	.005	.040	.004	.056	.006
			10	.090	.010	.054	.006	.040	.005	.039	.004	.053	.006
			30	.064	.008	.041	.006	.030	.004	.029	.003	.040	.005
1000	300	N(0,1)	5	.147	.020	.068	.009	.064	.008	.081	.010	.128	.022
			10	.141	.019	.064	.007	.060	.007	.077	.010	.116	.018
			30	.078	.011	.046	.006	.037	.005	.038	.005	.054	.007
		N(1,1)	5	.140	.018	.081	.009	.062	.007	.058	.006	.083	.010
			10	.134	.017	.077	.009	.059	.007	.056	.006	.077	.009
			30	.079	.011	.050	.007	.037	.006	.036	.004	.051	.007

[a] Base ability is N(1,1).

[b] Number of Common Items

Table 6

Mean and Standard Deviation of Mean Euclidean Distances over 100 Replications

Sample Size		Target		Separate Calibration		Concurrent Calibration	
Base	Target	Ability[a]	NC[b]	Mean	SD	Mean	SD
300	300	N(0,1)	5	.255	.018	.249	.019
			10	.241	.016	.236	.015
			30	.184	.013	.167	.011
		N(1,1)	5	.233	.016	.228	.014
			10	.221	.014	.219	.014
			30	.166	.013	.164	.012
1000	1000	N(0,1)	5	.140	.012	.137	.009
			10	.131	.010	.129	.009
			30	.097	.008	.090	.007
		N(1,1)	5	.127	.009	.124	.008
			10	.120	.008	.119	.008
			30	.088	.007	.088	.007
1000	300	N(0,1)	5	.206	.016	.197	.015
			10	.195	.014	.185	.014
			30	.152	.013	.109	.009
		N(1,1)	5	.183	.013	.176	.012
			10	.173	.012	.166	.012
			30	.133	.010	.110	.008

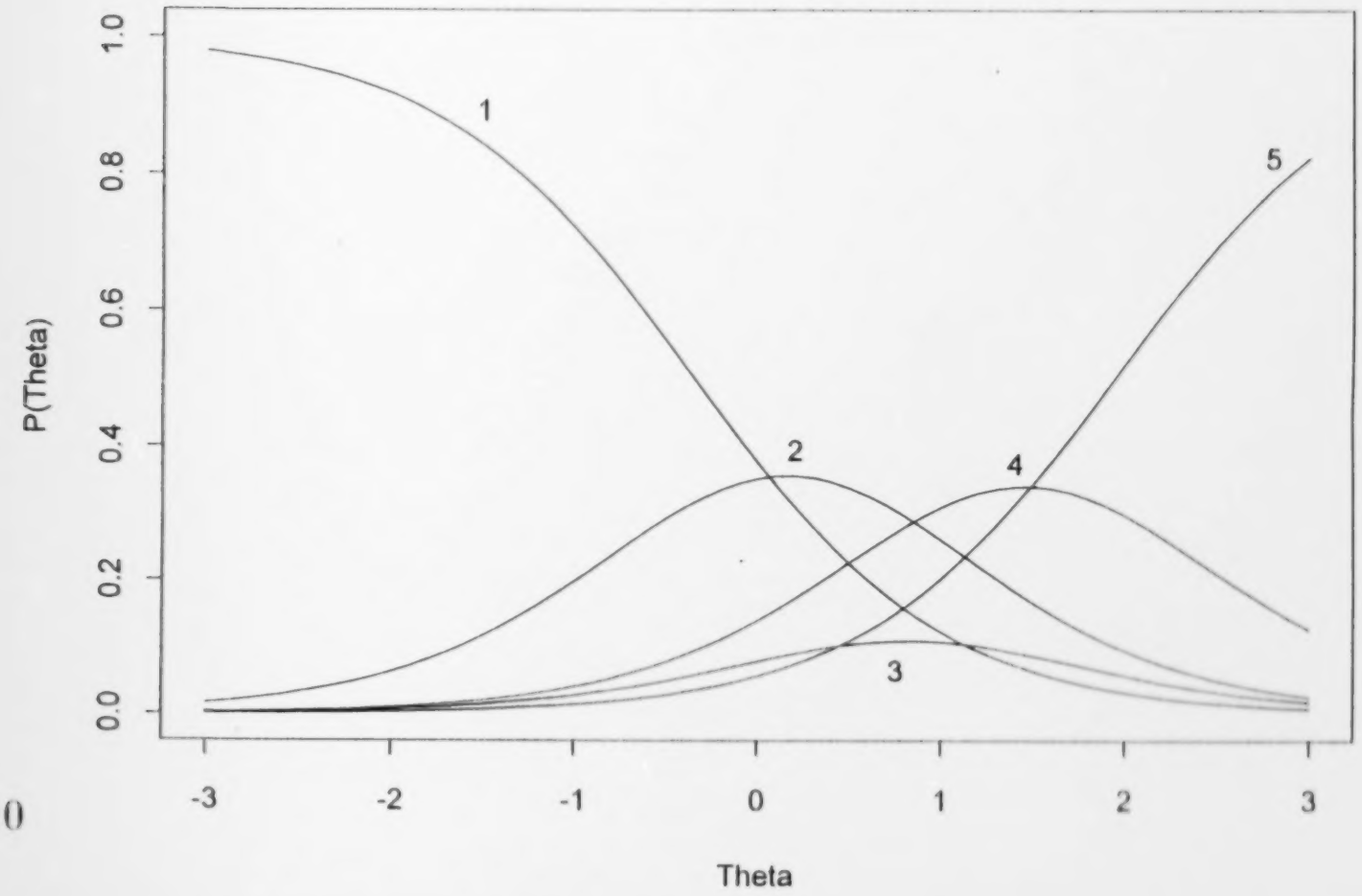
[a]Base ability is N(1,1).

[b]Number of Common Items

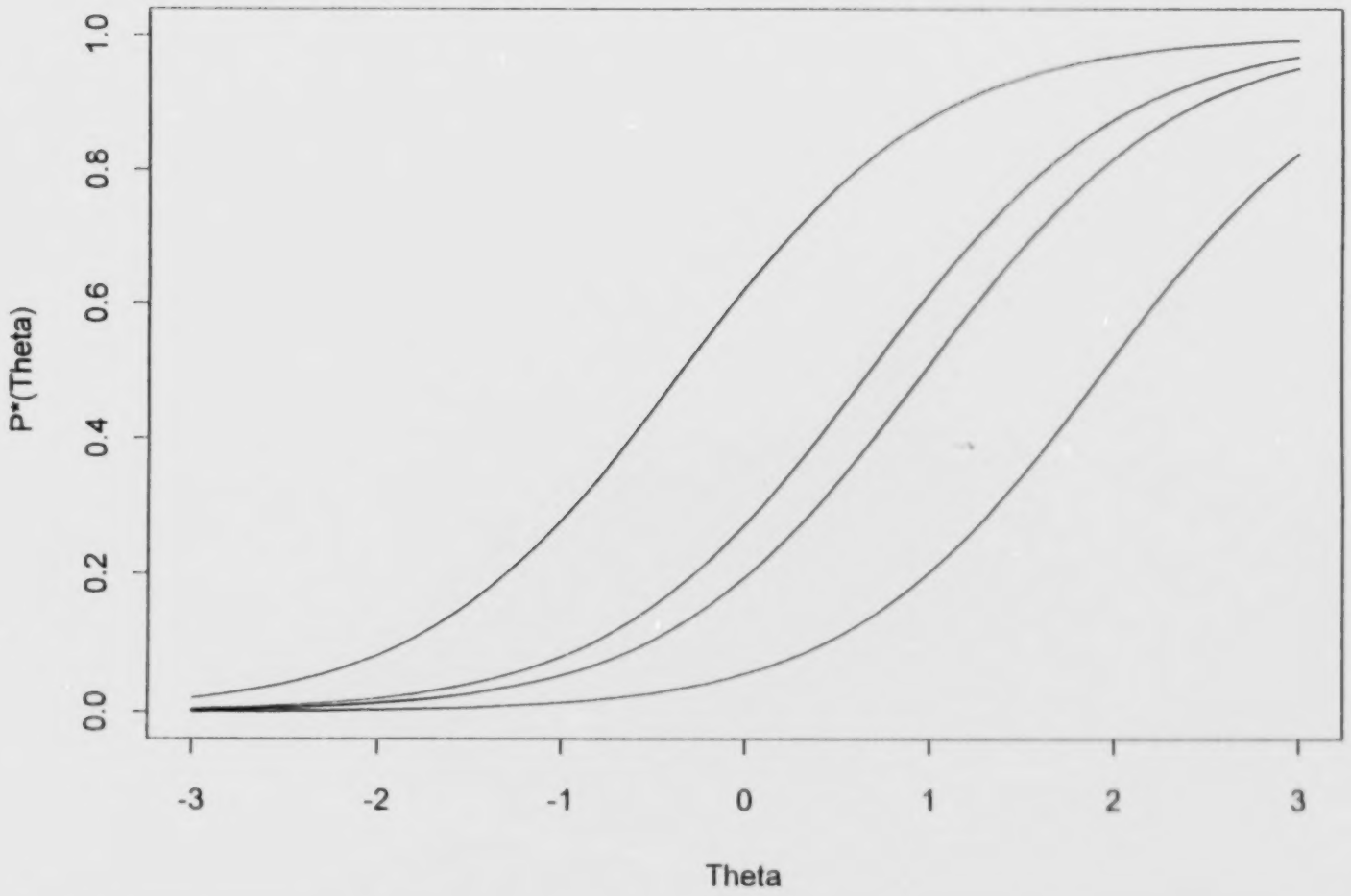
Figure Captions

Figure 1. Category Response Functions for a Five-Category Item

Figure 2. Boundary Response Functions for a Five-Category Item


Category Response Functions for a Five-Category Item
P(Theta)
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
-3
-2
-1
0
1
2
3
Theta

Boundary Response Functions for a Five-Category Item



Acknowledgments

The authors express their gratitude to Frank B. Baker for making his EQUATE and GENIRV programs available for this study.

Authors' Addresses

Send requests for reprints or further information to Seock-Ho Kim, The University of Georgia, 325 Aderhold Hall, Athens GA 30602, U.S.A., or Allan S. Cohen, Testing and Evaluation Services, University of Wisconsin, 1025 West Johnson Street, Madison WI 53706, U.S.A. Internet: skim@coe.uga.edu or cohen@tne.edsci.wisc.edu